SSOAR

Open Access Repository
www.ssoar.info

A Comparison of Two Cognitive Pretesting Techniques Supported by Eye Tracking

Neuert, Cornelia; Lenzner, Timo

Postprint / Postprint
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Neuert, C., & Lenzner, T. (2016). A Comparison of Two Cognitive Pretesting Techniques Supported by Eye Tracking. *Social Science Computer Review*, *34*(5), 582-596. https://doi.org/10.1177/0894439315596157

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-58069-9

This is an accepted manuscript of an article published in
Social Science Computer Review, Vol. 34, 2016, Iss. 5, pp. 582-596
Copyright © 2015 (The Authors). Reprinted by permission of SAGE Publications
DOI: 10.1177/0894439315596157
Page numbers have been adjusted to the publishers version, whereby this postprint is fully quotable

A Comparison of Two Cognitive Pretesting Techniques Supported by Eye Tracking

Cornelia E. Neuert and Timo Lenzner (GESIS—Leibniz Institute for the Social Sciences, Mannheim, Germany)

Corresponding Author:
Cornelia E. Neuert, GESIS—Leibniz Institute for the Social Sciences, Mannheim, B2, I, Mannheim D-68159, Germany.
Email: cornelia.neuert@gesis.org

Abstract

In questionnaire pretesting, supplementing cognitive interviewing with eye tracking is a promising new method that provides additional insights into respondents' cognitive processes while answering survey questions. When incorporating eye tracking into cognitive interviewing, two retrospective probing techniques seem to be particularly useful. In the first technique —retrospective probing— participants complete an online questionnaire, while cognitive interviewers monitor participants' eye movements in an adjacent room and note down any peculiarities in their reading patterns. Afterward, the interviewers ask targeted probing questions about these peculiarities in a subsequent cognitive interview. In the second technique—gaze video cued retrospective probing—respondents are additionally shown a video of their eye movements during the cognitive interview. This video stimulus is supposed to serve as a visual cue that may better enable respondents to remember their thoughts while answering the questions. We examine whether one of the two techniques is more effective when it comes to identifying problematic survey questions. In a lab experiment, participants' eye movements (n = 42) were tracked while they completed six questions of an online questionnaire. Simultaneously, their reading patterns were monitored by an interviewer for evidence of response problems. After completion of the online survey, a cognitive interview was conducted. In the retrospective probing condition, probing questions were asked if peculiar reading patterns were observed during the eye-tracking session (e.g., rereadings of specific words or text passages). In the other condition, participants were shown a video of their recorded eye movements, in addition to receiving probing questions about the questions displayed. Results show that both techniques did not differ in terms of the total number of problems identified. However, gaze video cued retrospective probing identified fewer unique problems and fewer types of problems than pure retrospective probing.

Keywords

cognitive interviews, eye tracking, pretesting, gaze-replay videos, web surveys

Introduction

The general goal of cognitive interviewing is to obtain information about the cognitive processes underlying survey responding and to identify difficulties respondents have in answering them. By identifying problematic questions and providing information about how a question could be revised, cognitive interviewing contributes to a better understanding of questions by respondents and thus decreases measurement error (Forsyth & Lessler, 1991; Willis, 2005). For example, measurement error is introduced into the data if respondents misinterpret words, concepts, or entire questions, have difficulties in retrieving the information sought, or encounter problems when formatting their answers (Groves et al., 2004, p. 209).

In questionnaire pretesting, supplementing cognitive interviewing with eye tracking is a novel and promising approach that might provide additional insights into respondents' cognitive processes while answering survey questions (Galesic & Yan, 2011). Whereas cognitive interviews initially took place in pretesting laboratories equipped with video and audio recording equipment, these labs are, today, often additionally equipped with eye-tracking technology (Campanelli, 2008); for instance, those at the German Federal Statistical Office (Tries, 2010) and at the U.S. Census Bureau (Romano & Chen, 2011). Incorporating eye tracking into cognitive interviewing is based on the idea of a direct relationship between eye movements and cognitive processing. The so-called eye-mind hypothesis of Just and Carpenter (1980) assumes a link between what people are looking at and what they are thinking. It postulates that words or objects are fixated as long as they are being processed (Just & Carpenter, 1980). According to this assumption, eye tracking appears to be a natural supplement to cognitive interviewing because cognitive interviewing is about obtaining information about people's thoughts while answering a questionnaire (Willis, 2005). Observing the eye movements—where and for how long respondents look when reading and answering questions—helps to reach a better understanding of the participant's answer process and can be used to detect difficulties that may have arisen while answering (Neuert & Lenzner, 2015). Because eye tracking allows the detection of conscious and unconscious reactions to survey questions (Tries, Nebel, & Blanke, 2012), it might also point to difficulties that are not consciously apparent to participants and have a small chance of being detected (Blair & Conrad, 2011). As we have demonstrated in a previous study, incorporating eye tracking into cognitive interviewing is indeed more productive in identifying questionnaire problems than using cognitive interviewing alone (Neuert & Lenzner, 2015).

In the present article, we are interested in how eye tracking can be implemented most effectively into cognitive survey pretesting studies. We compare two eye tracking supported cognitive pretesting techniques: Retrospective probing based on observed eye movements and retrospective probing, which incorporates a gaze video cue, that is, a video that shows the participants' eye movements while they filled in an online questionnaire.

Background

The term ‘‘cognitive interviewing” usually refers to administering draft questions of a survey instrument to respondents who provide additional verbal information about their responses and their thoughts (Beatty & Willis, 2007). Cognitive interviewing aims to understand and to obtain information on respondents' thought processes while answering these questions (i.e., how respondents understand the questions, as well as how they arrive at an answer) and to identify specific difficulties respondents have with the questionnaire (Beatty, 2004; Beatty & Willis, 2007). The verbal material about respondents' thought processes that is gathered in the cognitive interviews is used to evaluate the quality of the questions and to provide information about whether a question needs revision (Beatty & Willis, 2007).

One of the most common techniques used in cognitive interviews is ‘‘verbal probing.'' Probes are follow-up questions about what respondents were thinking and how they interpreted the questions or specific terms used in the questionnaire (Willis, 2005). During cognitive interviews, participants typically first answer the survey questions and then respond to a series of probing questions (Willis, 2005; Willis & Miller, 2011). Follow-up probing can occur either immediately after the subject has answered the target survey question (concurrent probing) or at the end of the interview, during a debriefing session (retrospective probing; Willis, 2005). In current practice, concurrent probing is used more frequently, although, under certain circumstances, retrospective probing may be the more efficient technique, for example, when testing self-administered questionnaires, in which the respondent should not be disturbed, to determine whether he or she can handle the instrument alone (Willis, 2005).

When conducting cognitive interviews in combination with eye tracking, it is sensible to probe only retrospectively. In eye-tracking supported cognitive pretesting studies, respondents are seated in front of an eye tracker in the laboratory and are instructed to fill in a questionnaire at their usual pace. Simultaneously, a cognitive interviewer monitors the respondents' actions and eye movements, in real time, on a computer screen in an adjacent room and notes any peculiarities in their reading patterns (e.g., long or repeated fixations or multiple regressions from answers to question text). These are then addressed in a cognitive interview that is conducted after respondents have completed the survey. If eye tracking were to be used with concurrent probing, participants might produce eye movements that they would not normally make when they complete an online questionnaire on their own (Pernice & Nielsen, 2009). For example, unusual eye movements might be caused by participants looking away from the screen when describing something to the interviewer or by fixating on certain areas of the screen while describing their thought processes regarding that question. Unusual eye movements would be especially disadvantageous if the data were also evaluated quantitatively after the interview. Concurrent probing might also make participants more aware of the fact that their eye movements are being tracked. Therefore, when conducting cognitive interviews in combination with eye tracking, it is reasonable to apply retrospective rather than concurrent probing.

In general, retrospective probing has the advantage that it does not interrupt the flow of answering an entire questionnaire and, thus, creates a more realistic field setting. However, retrospective probing also has some drawbacks, because participants may have forgotten key information or the information about their problems may no longer be accessible when they are finally asked to answer the probing questions (Willis, 2005). A potential solution to aid the participants' memory could be the use of a gaze video cue, a technique that has already been employed in usability research in combination with thinking aloud (e.g., Ball, Eger, Stevens, & Dodd, 2006; Elling, Lentz, & DeJong, 2011; Hansen, 1991; Hyrskykari, Ovaska, Majaranta, Raiha, & Lehtinen, 2008) as well as in field research with mobile eye tracking (Eghbal-Azar & Widlok, 2013). When using retrospective probing in conjunction with a gaze video, participants are presented with a replay of their eye movements during the cognitive interview. In the video replay, the eye movements appear as red dots that represent where participants were looking when answering the questions. The longer a participant looks at something, the larger the red dot becomes. Thus, it is possible for the participant to see how he or she read and answered the question. This video stimulus is supposed to serve as a visual cue that may better enable respondents to remember their thoughts while answering the questions by reviewing their eye movements.

On the negative side, showing participants a gaze video replay may increase the risk of false alarms, that is, identifying a problem that is not actually present (Conrad & Blair, 2009). When confronted with their own eye movements, participants might come up with a post hoc explanation for their behavior to meet what they think is expected of them, instead of just reporting their thinking.

In this study, we compare gaze video cued retrospective probing with retrospective probing without any cues within the framework of identifying problematic survey questions. Three research questions will be addressed:

Research question 1: Do both techniques differ in terms of the number of problems identified?

Research question 2: Do both techniques differ in the types of problems identified?

Research question 3: Do both techniques differ in the way they stimulate participants when commenting on their behavior?

Methods

Design

To answer our research questions, we used a randomized between-subject design with two conditions (gaze-replay video yes/no). All participants (n = 42) were seated in front of the eye tracker and, after a short explanation of the eye tracker and a standard calibration procedure, the participants completed the online questionnaire while their eye movements were recorded and their response behavior was monitored by a cognitive interviewer sitting in a different room. The interviewer used a coding scheme (described in Interview protocol & interviewer instructions section) to document any peculiar reading pattern that was observed. Following completion of the online survey, a cognitive interview was conducted. Each cognitive interview was videotaped. During the cognitive interview, participants in the retrospective probing condition (n = 21) received a paper version of the questionnaire with screenshots of the questions, to remind them of their initial thoughts, whereas participants in the gaze video cued retrospective probing condition (n = 21) were shown a video of their recorded eye movements while filling in the online questionnaire. In addition, respondents in both conditions were asked a set of probing questions about the questions under scrutiny.

Participants

This experiment was part of a larger study conducted in October and November 2012 in the pretest laboratory at GESIS—Leibniz Institute for the Social Sciences in Mannheim, Germany (see Procedure section for detailed information). For this experiment, 33 participants were recruited from the respondent pool maintained by the institute, as well as by word of mouth. For their participation in the whole study, which took about 1 ½ hr, participants received a compensation of €30. Additionally, nine colleagues and student assistants working primarily in nonscientific departments of the institute participated in the study for free, so that a total of 42 subjects participated in the experiment. Participants came separately to the pretest laboratory at GESIS for individual sessions. Table 1 shows some demographic characteristics of the participants.

Questionnaire

The questionnaire included 6 closed-ended items that were adapted from the International Social Survey Programme (2003, 2004) and the European Social Survey (round 1, 2002; round 5, 2010). The language of the questionnaire was German. The official English translations of the questions provided by the survey organizers are available in Appendix A. The questions included two question formats: four single-choice questions and one grid question with 2 items. One of the questions asked

Table 1. Demographic Characteristics of Participants (%).

Gender		Age		Years of Schooling		Computer Usage	
Female	52%	18–34	60%	9 years or less	19%	(Almost) Daily	91%
Male	48%	35–54	33%	10 years	10%	Weekly	2%
		55+	7%	12 years or more	71%	Seldom or never	7%

about respondents' behavior, the other five about respondents' attitudes. The online questionnaire was programmed with a font size of 18 and 16 pixels and a line height of 40 and 32 pixels for the question text and answer options, respectively.

Eye-Tracking Equipment

We used a Tobii T120 eye-tracking system together with the Tobii Studio 3.2.1 software to record the participants' eye movements. The Tobii T120 is a remote eye tracker embedded in a 17" TFT monitor (resolution 1,280 x 1,024) with two binocular infrared cameras placed underneath the computer screen. This system is particularly suitable when stimuli can be presented on a screen and provides unobtrusive recording of respondents' eye movements and permits head movements within a scale of 30 x 22 x 30 cm. Eye movements were recorded at a sampling rate of 120 Hz, meaning that 120 gaze data points per second were collected for each eye. The Tobii Studio software allows the interviewer to play back a video recording of the original recording, with or without eye movements; in our case, a video of the respondents' eye movements recorded during completion of the online questionnaire. The software also includes an automatic retrospective think-aloud recording function that allows the interviewer to video and audio record the participants' comments and reactions while showing a playback from the previously recorded task. Finally, the software includes features that enable the interviewer to adjust playback speed, start or pause playing, rewind or fast forward the video. This allows the interviewer to control the recording, for example, to pause if the participant needs more time to respond, or to repeat a video sequence.

Interview Protocol and Interviewer Instructions

The interview protocol included prescripted, general probing questions for all 6 items, such as ‘‘Could you please explain your answer a little further?" ‘‘What were you thinking when answering the question?" ‘‘How easy or difficult was it for you to come up with your answer?" and ‘‘Why did you find it (rather/very) difficult?" The use of prescripted probing questions ensured a relatively standardized application of the protocol between the different interviewers. The use of general probing (in contrast to specific probing) questions has the advantage that they do not influence the answer process of the respondent. Furthermore, general probes induce the participant to elaborate in a narrative way, which helps to collect information on how and why respondents answered the question as they did (Willson & Miller, 2014).

The interviewers were instructed to probe only those questions for which peculiar reading patterns were observed during the

eye-tracking session. To document if a peculiarity occurred, interviewers were provided with a coding scheme for peculiar reading patterns: They had to check a box if they observed one of the following five behaviors: (1) long or repeated fixations on a word, (2) rereadings of specific words or text passages, (3) regressions from answers to question text, (4) correction of the chosen response category, and (5) skipping a question. In addition, it was possible to check a box if an ‘‘other,'' not specified peculiarity occurred and to describe the corresponding behavior. If one or more of the behaviors described previously were observed during the eye-tracking session, the interviewers were instructed to first ask the general probing questions and to probe the peculiar reading patterns explicitly only if the general probes had not already uncovered the reasons for this particular behavior.

Participants in the gaze video cued retrospective probing condition were given the following instruction: ‘‘I am now going to show you a recording of your eye movements during/while answering question x. The red dots that you are going to see in the replay show how you read and answered the question and represent where you were looking. The longer you were looking at something, the larger the red dot becomes. After you have watched the replay, I would like you to tell me how you came up with your answer and what you were thinking when answering the question.''

Procedure

The experiment reported in this article was part of al arger study with several unrelated experiments. The entire study took about 1 ½ hr and consisted of three parts. In the first part, participants completed an online questionnaire while their eye movements were tracked. The entire questionnaire included 58 questions. In the second part, a cognitive interview was conducted (cf. Neuert & Lenzner, 2015). In the third part, participants completed another online questionnaire that consisted of different small experiments unrelated to this study (cf. Lenzner, Kaczmirek, & Galesic, 2014). The experiment reported in this article refers to the last six questions of the online questionnaire (part one of the study), which were discussed at the end of the subsequent cognitive interview (part two of the study). The interviews in both conditions were conducted by five interviewers (three researchers and two student assistants) who had all previously conducted cognitive interviews. Individual interviewers each conducted between three to five interviews in each condition. The average survey completion time for the six questions was approximately 2.5 min (154 s). In terms of time required for conducting the cognitive interviews in both conditions, we found that administering retrospective probing in conjunction with a gaze video cue required close to 373 s, whereas the pure retrospective probing interviews took approximately 331 s.

Results

In the analysis described subsequently, we compared gaze video cued retrospective probing and retrospective probing both quantitatively, that is, in terms of the total number of problems identified (including recurrences of the same problem) and the number of unique problems identified, and qualitatively, that is, in terms of the types of problems identified and the types of comments given by respondents. First, we examined the total number of problems identified in each condition. Subsequently, we categorized the types of problems and examined the number of unique problems. Finally, we categorized the types of comments given by respondents.

Number of Problems

To identify problems, the first author reviewed all videotapes of the cognitive interviews and gave each questionnaire item, for each interview, a dichotomous score that reflected whether a problem was identified in the question (1) or not (0). Those sections of the cognitive interviews that contained a context relevant for understanding potential problems were transcribed. Afterward, a student assistant reviewed and coded all interviews, to estimate interrater reliability. Agreement between these two raters was 93% and Cohen's Kappa (1960) was found to be .84, which is ‘‘almost perfect,'' according to Landis and Koch's (1977, p. 165) criteria. The number of problems that resulted from this analysis

Table 2. Number of Problems Identified, by Condition.

		Number of Problems in					
	Total Number of Problems	Question 1.1	Question 1.2	Question 2	Question 3	Question 4	Question 5
Retrospective probing (no cue)	41	1 (2%)	10 (24%)	7 (17%)	8 (20%)	5 (12%)	10 (24%)
Gaze video cued retrospective probing (video cue)	44	1 (2%)	9 (21%)	5 (11%)	9 (21%)	7 (16%)	13 (30%)

Table 3. Types of Problems Identified, by Condition.

		Types of Problems			
	Total Number of Problems	Comprehension	Retrieval	Judgment	Response Selection
Retrospective probing (no cue)	41	36 (88%)	0 (0%)	3 (7%)	2 (5%)
Gaze video cued retrospective probing (video cue)	44	40 (91%)	0 (0%)	4 (9%)	0 (0%)

contained all detected problems for all participants, which means that problems can occur repeatedly for specific questions, because several participants might have encountered the same problem.

Table 2 shows the total number of problems identified in each condition and the distribution of these problems per question. A comparison of the total number of problems across conditions revealed that the combination of a gaze video with retrospective probing did not identify significantly more problems (n = 44) than retrospective probing (n = 41; $x^2 = 1.38$, df = 1, p = .160).

In both conditions, most problems were identified in Question 5 (23 problems) and in Question 1.2 (19 problems), whereas only one participant in each condition experienced a problem when answering Question 1.1.

Types of Problems

In our next analysis step, we evaluated whether both techniques identified different types of problems. For each item that was perceived as problematic, we reviewed the transcripts of the interviews and coded them into problem types, using a problem classification scheme adopted from various existing schemes (DeMaio & Landreth, 2004; Lessler & Forsyth, 1996; Presser & Blair, 1994; Rothgeb, Willis, & Forsyth, 2001).

The problem classification scheme included a total of 30 problem codes that were grouped according to the four stages of the survey response process (comprehension, retrieval, judgment, and response selection; Tourangeau, 1984; Tourangeau, Rips, & Rasinski, 2000; see Appendix B). Individual items could be assigned to multiple problem codes. Problem types were also coded by a student assistant, resulting in an agreement of 79% and a k of.74 (classified as ‘‘substantial" reliability by Landis & Koch, 1977). The types of problems discovered in the questions came from three of the four stages of the survey response process: comprehension difficulties, judgmental issues, and response selection. Problems with information retrieval were not detected (see Table 3).

Table 4. Number of Unique Problems Identified, by Condition.

		Number of Problems in					
	Total Number of Problems	Question 1.1	Question 1.2	Question 2	Question 3	Question 4	Question 5
Retrospective probing (no cue)	20	1 (5%)	3 (15%)	3 (15%)	4 (20%)	5 (25%)	4 (20%)
Gaze video cued retrospective probing (video cue)	14	1 (7%)	1 (7%)	2 (14%)	3 (21%)	4 (29%)	3 (21%)

In both conditions, the highest proportion of problems was classified as comprehension problems. Two types of problems from the ‘‘response selection" category were detected in the retrospective probing condition, but problems with response selection were not found in the gaze video cue retrospective probing condition. Again, no statistically significant difference between the two conditions, with regard to the types of problems identified, was found ($x^2 = 2.25$, df = 2, p = .325).

Besides the general productivity of each technique, it is important to establish how many unique problems each technique identified. We therefore also looked at the number of unique problems detected in each condition (Table 4). We classified a problem as unique if it occurred at least once per question (irrespective of how many participants had experienced the same problem). When comparing the total number of unique problems across conditions, we found that gaze video cued retrospective probing identified significantly less unique problems (n = 14) than retrospective probing (n = 20; $x^2 = 5.56$, df = 1, p = .037).

Although, in Question 1.1, one problem in the retrospective probing condition and one in the gaze video cued retrospective probing condition were detected, retrospective probing identified one (Questions 2-5) or even two unique problems (Question 1.2) more than gaze video cued retrospective probing had detected in all other questions. Whereas, in Question 1.2, the problem that the term ‘‘civil disobedience" was unknown to some respondents (Code 4, see Table 5) was identified in both conditions, two other problems were identified exclusively in the retrospective probing condition. In this condition, the question was also found to be vague and unclear (Code 1) and to have a complex syntactical structure (Code 11). Altogether, three unique problems were detected in Question 2. Even though two problem types, namely, that the question was vague and unclear (Code 1) and that it contained a complex topic (Code 2), were identified in both conditions, the more specific problem—the question contained undefined terms (United Nations; intervene)—was only detected in the retrospective probing condition. A summary of the number and types of problems identified per question and condition is presented in Table 5.

In both conditions, the highest proportion of unique problems was classified as ‘‘vague or unclear question" (25% retrospective probing and 29% gaze video cued retrospective probing), or as containing ‘‘undefined or vague terms" (20% retrospective probing and 21% gaze video cued retrospective probing). Four types of unique problems were detected exclusively in the pure retrospective probing condition: Only respondents in this condition referred to the error codes ‘‘knowledge may not exist" (Question 4), ‘‘erroneous or inappropriate assumption" (Question 3), ‘‘response categories missing" (Question 5), and ‘‘no formally adequate answer" (Question 4).

Classification of Comments

To examine whether the different cues stimulate the participants in different ways when commenting on their behavior, we classified participants' comments into three categories, according to the

Table 5. Number and Types of Unique Problems Identified, by Condition.

Questions		Number of Unique Problems	Types of Problems (Code)	Frequency
Question 1.1	No video	1	Undefined/vague term (opportunities to participate in public decision-making) (4)	1
	Video cued	1	Undefined/vague term (opportunities to participate in public decision-making) (4)	1
Question 1.2	No video	3	Undefined/vague term (civil disobedience) (4)	8
			Vague/unclear question (1)	1
			Complex or awkward syntax (11)	1
	Video cued	1	Undefined/vague term (civil disobedience) (4)	9
Question 2	No video	3	Vague/unclear question (1)	1
			Complex topic (2)	2
			Undefined/vague term (United Nations; intervene) (4)	4
	Video cued	2	Vague/unclear question (1)	1
			Complex topic (2)	4
Question 3	No video	4	Vague/unclear question (1)	2
			Complex or awkward syntax (11)	3
			Potentially sensitive or desirability bias (21)	1
			Erroneous/inappropriate assumption (12)	2
	Video cued	3	Vague/unclear question (1)	2
			Complex or awkward syntax (11)	6
			Potentially sensitive or desirability bias (21)	1
Question 4	No video	5	Vague/unclear question (1)	1
			Complex topic (2)	1
			Undefined/vague term (direct influence) (4)	1
			Knowledge may not exist (5)	1
			No formally adequate answer (28)	1
	Video cued	4	Vague/unclear question (1)	2
			Complex topic (2)	1
			Undefined/vague term (direct influence) (4)	3
			Boundary lines (6)	1
Question 5	No video	4	Vague/unclear question (1)	1
			Boundary lines (6)	6
			Complex estimation (20)	2
			Response categories missing (27)	1
	Video cued	3	Vague/unclear question (1)	4
			Boundary lines (6)	6
			Complex estimation (20)	3

coding scheme of verbalizations suggested by Hansen (1991), which was slightly altered for our purposes (see Table 6). Instead of speaking of ‘‘manipulative operations" that describe an action in a usability test (Hansen, 1991), we used the term ‘‘behavioral" to code comments that express exclusively an action, for example ‘‘I have read the question and answered it." ‘‘Cognitive" comments are defined as interpretations, assessments, and expectations of the respondents (e.g., I have never heard the term [x] before). Our third category is a combination of both, where ‘‘cognitive and behavioral" comments are associated with each other, for example ‘‘I wasn't sure about the term [x] and that is why I read the question several times." For the classification of comments, we coded all those sections of the cognitive interviews that contained a relevant context for understanding whether a problem existed or not. A total of 95 comments (48 in the retrospective probing condition and 47 in the

Table 6. Class of Comments, by Condition.

	Total Number of Comments	Types of Comments		
		Behavioral	Cognitive	Behavioral–Cognitive
Retrospective probing (no cue)	48	2 (4%)	31 (65%)	15 (31%)
Gaze video cued retrospective probing (video cue)	47	5 (11%)	25 (53%)	17 (36%)
Total	95	7 (7%)	56 (59%)	32 (34%)

gaze video cued retrospective probing condition, see Table 6) were coded by the first author and a student assistant, respectively. Interrater reliability between both coders was found to be k = .78, which is generally classified as “substantial” reliability (Landis & Koch, 1977, p. 165) and agreement was found to be 87%. Only one code was assigned to each comment. The results are shown in Table 6. With respect to the types of comments, gaze video cued retrospective probing stimulated the participants to produce slightly more ‘‘behavioral" comments (11% vs. 4%) and to produce less ‘‘cognitive" comments than when no cue was used (53% vs. 65%), meaning that participants were commenting more on what they were doing and less on what they were thinking when answering questions.

The gaze video cued retrospective probing condition also stimulated the participants to produce slightly more ‘‘behavioral and cognitive" comments (36% vs. 31%) in which the participants linked their behavior with what they were thinking at the time. Overall, the highest proportion of comments was classified as ‘‘cognitive" in both conditions.

In order to evaluate how well the technique of gaze video cued probing worked, we took brief notes after reviewing each cognitive interview in the gaze video cued probing condition and categorized participants into three groups: technique worked well, moderately well, or not at all. For almost half of the participants (n = 9), seeing a replay of their own eye movements worked well and they were able to associate what they were seeing with what they had been thinking. For a further eight participants, the technique worked moderately well. However, in this group, after a period of adaptation, the technique worked increasingly better toward the end of the interview. The remaining four participants had problems with the task and were either simply looking at their eye movements or were describing what they were seeing, but not referring to the question.

Discussion and Conclusion

The goal of this experiment was to compare retrospective probing, in conjunction with a gaze video replay, with retrospective probing without any cue when testing survey questions in pretesting studies supported by eye tracking. Results show that the combination of retrospective probing with a gaze video cue and the pure retrospective probing did not differ significantly in terms of their quantitative output (i.e., total number of problems identified). However, gaze video cued retrospective probing identified significantly fewer unique problems and fewer types of problems. Hence, we do not find evidence that eye movement replay serves as an extra cue that enables participants to better remember what they were thinking when answering the questions. However, due to the relatively small sample size of this study, our conclusions have to be considered with caution and we encourage further methodological investigations to confirm or reject our results.

A potential explanation for why the gaze video cue did not produce better results than pure retrospective probing might be that the eye movements not only supported participants in remembering their initial thoughts, but also distracted them. For most participants, seeing their own eye movements was a new experience. Although we explained to them what they would see, we observed that it was often difficult for participants to interpret the replay of their eye movements. The categorization of the comments made by the participants revealed that gaze video cued retrospective probing stimulated the participants to produce slightly more ‘‘behavioral'' comments and to produce fewer ‘‘cognitive'' comments than when no cue was used. Seeing a replay of their own eye movements might have stimulated the participants simply to describe what they were doing instead of what they were thinking while answering the questions. In line with this argument, by exclusively describing what they were seeing, the participants might not have provided the interviewers with enough information to diagnose whether a problem existed and, if so, what caused the problem. In addition, we were concerned that the gaze video cue might increase the risk of false alarms, because participants could be tempted to provide post hoc explanations for their viewing behavior. However, our findings do not indicate that showing a gaze replay increased the risk of false alarms. Even though gaze video cued retrospective probing identified slightly more problems than pure retrospective probing, both techniques did not differ in the types of identified problems and retrospective probing identified even more unique problems than video cued retrospective probing.

Our results are limited by a number of factors that encourage additional studies. First, the cognitive interviewing protocol was prescripted and relatively structured, so that interviewers were not encouraged to probe spontaneously. Furthermore, we exclusively asked general probing questions and did not use specific probes (specially designed to address response processes within the four-stage cognitive model). In cognitive interviews, interviewers typically probe participants' responses in a more flexible manner and it might be worth examining whether more specific questions that are based on the observed eye movements have a positive effect on respondents remembering what they thought while seeing their eye movements. Maybe we would have identified more, or other, problems if interviewers had been given more flexibility, which is a general strength of cognitive interviewing as a pretesting method. Additionally, the experiment reported in this article was conducted only for the last six questions of a longer questionnaire and participants answered probing questions for the other questions without seeing a video of their eye movements in a previous part of the cognitive interview. By the time, the gaze video recording was shown, some respondents might have got used to the previously applied probing style and seeing the video recording of their eye movements in addition might have caused confusion. Furthermore, the benefit of the eye movement replay might have been stronger if participants had been given more time to habituate to the recording. Hence, it may be worth investigating whether training respondents in interpreting their eye movements for a few minutes before starting the actual interview and using the gaze video cue earlier in the cognitive interview could render the technique more useful.

Another limitation of our study is that we used relatively short survey questions. It is possible that the technique is not, or less, suitable for short survey questions or short texts in general. The added value of showing participants a video of their eye movements might be greater when websites or more complex question designs, such as those used in business surveys, are tested; these require an enhanced interaction with an online questionnaire or website (e.g., questions with lookup databases, question navigation with tabs). We encourage future research on questions in which more complex designs are used. For those questions, it might also be worth to compare whether seeing a replay of the answer process without the gaze overlay might

decrease participants confusion which could thus be more effective than seeing a video replay with a gaze overlay when identifying question problems. A final limitation is that no concurrent techniques such as thinking aloud or concurrent probing techniques were used in this experiment. Future research could investigate whether combining the gaze video cue with thinking aloud or concurrent probing might be more appropriate than combining it with retrospective verbal probing. With regard to the practical implications of this study, our findings suggest that using a gaze video replay in combination with retrospective probing is not worth the effort when pretesting short survey questions, because gaze video cued retrospective probing identified significantly less unique problems and less types of problems than pure retrospective probing. Moreover, the application of a gaze video replay is more time consuming than simple verbal probing and some participants clearly had difficulties in interpreting their own eye movements, which might have distracted them from reporting problems they had actually experienced when answering the questions. We therefore do not recommend the use of gaze video cued retrospective probing in eye tracking supported pretesting studies unless there is a special interest in usability and questionnaire navigation that should be discussed with participants.

Appendix A

Questions

Question 1. There are different opinions about people's rights in a democracy. On a scale of 1 to 7, where 1 is not at all important and 7 is very important, how important is it: Please tick one box on each line.
Question 1.1. That people be given more opportunities to participate in public decision-making.
Question 1.2. That citizens may engage in acts of civil disobedience when they oppose government actions.

Question 2. Which of these two statements comes closer to your view?
If a country seriously violates human rights, the United Nations should intervene.
Even if human rights are seriously violated the country's sovereignty must be respected, and the United Nations should not intervene.
Don't know what the United Nations is.

Question 3. How much do you agree or disagree with the following statements?
I am often less proud of Germany than I would like to be.
Agree strongly—Agree—Neither agree nor disagree—Disagree—Disagree strongly

Question 4. And how much would you say that the political system in [country] allows people like you to have a direct influence on politics? Please tick one box.
Not at all—Very little—Some—A lot—A great deal

Question 5. Not counting anything you do for your family, in your work, or within voluntary organizations, how often, if at all, do you actively provide help for other people?
Every day—Several times a week—Once a week—Several times a month—Once a month—Less often—Never

Appendix B

Table B1. Classification Scheme

Comprehension	Retrieval	Judgment	Response Selection
Question Content	*Retrieval from memory*	*Judgment and evaluation*	*Response terminology*
1. Vague/unclear question	18. High detail required or information unavailable	20. Complex estimation, difficult mental calculation required	22. Undefined/vague term
2. Complex topic	19. Long recall or reference period	21. Potentially sensitive/ desirability bias	*Response Units*
3. Topic carried over from earlier question			23. Response categories not appropriate to question
4. Undefined/vague term			24. Too detailed or broad response categories
5. Knowledge may not exist			25. Vague response categories
6. Boundary lines			*Response structure*
7. Objectively wrong answer, question is misunderstood			26. Overlapping response categories
Question structure			27. Response categories missing
8. Transition needed			28. No formally adequate answer
9. Unclear respondent instruction			29. Uncertainty which answer category reflects own opinion
10. Information overload, question too long			*Questionnaire Navigation*
11. Complex or awkward syntax			30. Questionnaire navigation
12. Erroneous/inappropriate assumption			
13. Assumes constant behavior			
14. Several questions in one, multiple subjects			
15. The response of others or of the general public is asked for			
Reference period			
16. Reference periods are missing or undefined			
17. Reference period carried over from earlier question			

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Ball, L., Eger, N., Stevens, R., & Dodd, J. (2006). Applying the post-experience eye-tracked protocol (PEEP) method in usability testing. Interfaces, 67, 15-19.

Beatty, P. C. (2004). The dynamics of cognitive interviewing. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer, Methods for testing and evaluating survey questionnaires (pp. 45-66). Hoboken, NJ: John Wiley.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. Public Opinion Quarterly, 71, 287-311. doi:10.1093/poq/nfm006

Blair, J., & Conrad, F. G. (2011). Sample size for cognitive interview pretesting. Public Opinion Quarterly, 75, 636-658. doi:10.1093/poq/nfr035

Campanelli, P. (2008). Testing survey questions. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), International handbook of survey methodology (pp. 176-200). New York/London: Erlbaum/Taylor & Francis.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. Public Opinion Quarterly, 73, 32-55.

DeMaio, T. J., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 89-108). Hoboken, NJ: John Wiley. doi:10.1002/0471654728.ch5

Eghbal-Azar, K., & Widlok, T. (2013). Potentials and limitations of mobile eye tracking in visitor studies: Evidence from field research at two museum exhibitions in Germany. Social Science Computer Review, 31, 103-118. doi:10.1177/0894439312453565

Elling, S., Lentz, L., & De Jong, M. (2011). Retrospective think-aloud method: Using eye movements as an extra cue for participants' verbalizations. In Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (pp. 1161-1170). New York, NY: ACM Press.

Forsyth, B. H., & Lessler, J. T. (1991). Cognitive laboratory methods: A taxonomy. In. P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), Measurement Errors in Surveys (pp. 393-418). New York, NY: John Wiley.

Galesic, M., & Yan, T. (2011). Use of eye tracking for studying survey response processes. In M. Das, P. Ester, & L. Kaczmirek (Eds.), Social and behavioral research and the internet: Advances in applied methods and research strategies (pp. 349-370). New York, NY: Routledge.

Groves, R. M., Fowler, F. J. Jr, Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). Survey methodology. Hoboken, NJ: John Wiley.

Hansen, J. P. (1991). The use of eye mark recordings to support verbal retrospection in software testing. Acta Psychologica, 76, 31-49.

Hyrskykari, A., Ovaska, S., Majaranta, P., Raiha, K.-J., & Lehtinen, M. (2008). Gaze path stimulation in retrospective think aloud. Journal of Eye Movement Research, 2, 1-18.

ISSP. (2003). International Social Survey Programme 2003: National Identity II (ISSP 2003). GESIS Data Archive, Cologne, Germany, ZA3910. Source Questionnaire.

ISSP. (2004). International Social Survey Programme 2004: Citizenship (ISSP 2004). GESIS Data Archive, Cologne, Germany, ZA3950. Source Questionnaire.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87, 329-354.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Lenzner, T., Kaczmirek, L., & Galesic, M. (2014). Left feels right: A usability study on the position of answer boxes in web surveys. Social Science Computer Review, 32, 743-764. doi:10.1177/0894439313517532

Lessler, J. T., & Forsyth, B. H. (1996). A coding system for appraising questionnaires. In N. Schwarz & S. Sudman (Eds.), Answering questions: Methodology for determining cognitive and communicative processes in survey research (pp. 259-291). San Francisco, CA: Jossey-Bass.

Neuert, C., & Lenzner, T. (2015). Incorporating Eye Tracking into Cognitive Interviewing to Pretest Survey Questions. International Journal of Social Research Methodology (online first). doi: http://dx.doi.org/10.1080/13645579.2015.1049448

Pernice, K., & Nielsen, J. (2009). How to conduct eyetracking studies. Fremont, CA: Nielsen Norman Group. Retrieved May 11, 2014, http://www.nngroup.com/reports/how-to-conduct-eyetracking-studies/

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results. Sociological methodology, 24, 73-104.

Romano, J. C., & Chen, J. M. (2011). A usability and eye-tracking evaluation of four versions of the online national survey of college graduates (NSCG): Iteration 2. Study Series: Survey Methodology 2011-01, Washington DC: U.S. Census Bureau.

Rothgeb, J., Willis, G., & Forsyth, B. (2001, May). Questionnaire pretesting methods: Do different techniques and different organizations produce similar results? Paper presented at the annual meeting of the American Association for Public Opinion Research, Montreal. Retrieved June 28, 2012, from https://www.census.gov/ srd/papers/pdf/rsm2005-02.pdf

Tourangeau, R. (1984). Cognitive science and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), Cognitive aspects of survey methodology: Building a bridge between disciplines (pp. 73-100). Washington, DC: National Academy Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. New York, NY: Cambridge University Press.

Tries, S. (2010). Usability tests of online questionnaires. In Federal Statistical Office (Ed.), Methods, approaches, developments: Information of the German federal statistical office (pp. 5-8). Wiesbaden, Germany: Federal Statistical Office.

Tries, S., Nebel, S., & Blanke, K. (2012). How to provide high data quality in online-questionnaires: Setting guidelines in design. Paper presented at the European Conference on Quality in Official Statistics, Athens, Greece, May 29-June 1, 2012. Retrieved November 19, 2014, from http://www.q2012.gr/articlefiles/sessions/34.1_Tries_On%20line%20questionnaires%20setting%20guidelines%20in%20design.pdf

Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design. Thousand Oaks, CA: Sage.

Willis, G. B., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. Field Methods, 23, 331-341.

Willson, S., & Miller, K. (2014). Data collection. In K. Miller, S. Willson, V. Chepp, & J. L. Padilla (Eds.), Cognitive interviewing methodology (pp. 15-33). Hoboken, NJ: John Wiley.

Author Biographies

Cornelia E. Neuert (cornelia.neuert@gesis.org) is a researcher at the Survey Design and Methodology Department at GESIS—Leibniz Institute for the Social Sciences, Germany. Her research interests include question evaluation, eye tracking, and survey methodology.

Timo Lenzner (timo.lenzner@gesis.org) is a senior researcher at the Survey Design and Methodology Department at GESIS—Leibniz Institute for the Social Sciences, Germany. His research focuses on questionnaire design and evaluation, Web surveys, and eye tracking.